Andrew Wiles' Marvellous Proof *

Henri Darmon (McGill University, Montreal, Canada)

Fermat famously claimed to have discovered "a truly marvellous proof" of his Last Theorem, which the margin of his copy of Diophantus' *Arithmetica* was too narrow to contain. While this proof (if it ever existed) is lost to posterity, Andrew Wiles' marvellous proof has been public for over two decades and has now earned him the Abel prize. According to the prize citation, Wiles merits this recognition "for his stunning proof of Fermat's Last Theorem by way of the modularity conjecture for semistable elliptic curves, opening a new era in number theory".

Few can remain insensitive to the allure of Fermat's Last Theorem, a riddle with roots in the mathematics of ancient Greece, simple enough to be understood and appreciated by a novice (like the 10-year-old Andrew Wiles browsing the shelves of his local public library), yet eluding the concerted efforts of the most brilliant minds for well over three centuries. It became, over its long history, the object of lucrative awards like the Wolfskehl prize and, more importantly, it motivated a cascade of fundamental discoveries: Fermat's method of infinite descent, Kummer's theory of ideals, the ABC conjecture, Frey's approach to ternary diophantine equations, Serre's conjecture on mod p Galois representations, ...

Even without its seemingly serendipitous connection to Fermat's Last Theorem, Wiles' modularity theorem is a fundamental statement about elliptic curves (as evidenced, for instance, by the key role it plays in the proof of Theorem 2 of Karl Rubin's contribution to the issue of the Notices of the AMS mentioned above). It is also a centrepiece of the "Langlands programme", the imposing, ambitious edifice of results and conjectures that has come to dominate the number theorist's view of the world. This programme has been described as a "grand unified theory" of mathematics. Taking a Norwegian perspective, it connects the objects that occur in the works of Niels Hendrik Abel, such as elliptic curves and their associated abelian integrals and Galois representations, with (frequently infinite-dimensional) linear representations of the continuous transformation groups, the study of which was pioneered by Sophus Lie. This report focuses on the role of Wiles' Theorem and its "marvellous proof" in the Langlands programme, in order to justify the closing phrase in the prize citation: how Wiles' proof has opened "a new era in number theory" and continues to have a profound and lasting impact on mathematics.

Our "beginner's tour" of the Langlands programme will only give a partial and undoubtedly biased glimpse of the full panorama, reflecting the author's shortcomings as well as the inherent limitations of a treatment aimed at a general readership. We will motivate the Langlands programme by starting with a discussion of *diophantine equations*: for the purposes of this exposition, they are equations of the form

$$\mathcal{X}: \quad P(x_1, \ldots, x_{n+1}) = 0, \tag{1}$$

where P is a polynomial in the variables $x_1, \ldots, x_{n+1}$ with integer (or sometimes rational) coefficients. One can examine the set, denoted $\mathcal{X}(F)$, of solutions of (1) with coordinates in any ring F. As we shall see, the subject draws much of its fascination from the deep and subtle ways in which the behaviours of different solution sets can resonate with each other, even if the sets $\mathcal{X}(\mathbb{Z})$ or $\mathcal{X}(\mathbb{Q})$ of integer and rational solutions are foremost in our minds. Examples of diophantine equations include Fermat's equation $x^d + y^d = z^d$ and the Brahmagupta-Pell equation $x^2 - Dy^2 = 1$ with $D > 0$, as well as elliptic curve equations of the form $y^2 = x^3 + ax + b$, in which a and b are rational parameters, the solutions (x, y) with rational coordinates being the object of interest in the latter case.

It can be instructive to approach a diophantine equation by first studying its solutions over *simpler* rings, such as the complete fields of real or complex numbers. The set

$$\mathbb{Z}/n\mathbb{Z} := \{0, 1, \ldots, n-1\} \tag{2}$$

of remainders after division by an integer $n \geq 2$, equipped with its natural laws of addition, subtraction and multiplication, is another particularly simple collection of numbers, of *finite cardinality*. If $n = p$ is *prime*, this ring is even a *field*: it comes equipped with an operation of division by non-zero elements, just like the more familiar collections of rational, real and complex numbers. The fact that $\mathbb{F}_p := \mathbb{Z}/p\mathbb{Z}$ is a field is an algebraic characterisation of the primes that forms the basis for most known efficient primality tests and factorisation algorithms. One of the great contributions of Evariste Galois, in addition to the eponymous theory that plays such a crucial role in Wiles' work, is his discovery of a field of cardinality p^r for any prime power p^r. This field, denoted $\mathbb{F}_{p^r}$ and sometimes referred to as the Galois field with p^r elements, is even *unique* up to isomorphism.

For a diophantine equation $\mathcal{X}$ as in (1), the most basic invariant of the set

$$\mathcal{X}(\mathbb{F}_{p^r}) := \{(x_1, \ldots, x_{n+1}) \in \mathbb{F}_{p^r}^{n+1} \text{ such that } P(x_1, \ldots, x_{n+1}) = 0\} \tag{3}$$

of solutions over $\mathbb{F}_{p^r}$ is of course its *cardinality*

$$N_{p^r} := \#\mathcal{X}(\mathbb{F}_{p^r}). \tag{4}$$

What patterns (if any) are satisfied by the sequence

$$N_p, N_{p^2}, N_{p^3}, \ldots, N_{p^r}, \ldots? \tag{5}$$

* This report is a very slightly expanded transcript of the Abel prize lecture delivered by the author on 25 May 2016 at the University of Oslo. It is published with the permission of the *Notices of the AMS*: reprinted from Volume 64, Issue 3, March 2017.

This sequence can be packaged into a generating series like

$$\sum_{r=1}^{\infty} N_{p^r} T^r \qquad \text{or} \qquad \sum_{r=1}^{\infty} \frac{N_{p^r}}{r} T^r. \tag{6}$$

For technical reasons, it is best to consider the exponential of the latter:

$$\zeta_p(\mathcal{X}; T) := \exp\left(\sum_{r=1}^{\infty} \frac{N_{p^r}}{r} T^r \right). \tag{7}$$

This power series in T is known as the *zeta function* of $\mathcal{X}$ over $\mathbb{F}_p$. It has integer coefficients and enjoys the following remarkable properties:

1. It is a *rational function* in T:
$$\zeta_p(\mathcal{X}; T) = \frac{Q(T)}{R(T)}, \tag{8}$$
where $Q(T)$ and $R(T)$ are polynomials in T whose degrees (for all but finitely many p) are *independent of* p and determined by the shape – the complex topology – of the set $\mathcal{X}(\mathbb{C})$ of complex solutions.
2. The reciprocal roots of $Q(T)$ and $R(T)$ are complex numbers of absolute value $p^{i/2}$ with i an integer in the interval $0 \leq i \leq 2n$.

The first statement – the rationality of the zeta function, which was proven by Bernard Dwork in the early 1960s – is a key part of the Weil conjectures, whose formulation in the 1940s unleashed a revolution in arithmetic geometry, driving the development of étale cohomology by Grothendieck and his school. The second statement, which asserts that the complex function $\zeta_p(\mathcal{X}; p^{-s})$ has its roots on the real lines $\Re(s) = i/2$ with i as above, is known as the Riemann hypothesis for the zeta functions of diophantine equations over finite fields. It was proven by Pierre Deligne in 1974 and is one of the major achievements for which he was awarded the Abel prize in 2013.

That the asymptotic behaviour of N_p can lead to deep insights into the behaviour of the associated diophantine equations is one of the key ideas behind the Birch and Swinnerton-Dyer conjecture. Understanding the patterns satisfied by the functions

$$p \mapsto N_p \qquad \text{and} \qquad p \mapsto \zeta_p(\mathcal{X}; T) \tag{9}$$

as the prime p varies will also serve as our motivating question for the Langlands programme.

It turns out to be fruitful to package the zeta functions over all the finite fields into a single function of a complex variable s, by taking the infinite product

$$\zeta(\mathcal{X}; s) = \prod_p \zeta_p(\mathcal{X}; p^{-s}) \tag{10}$$

as p ranges over all the prime numbers. In the case of the simplest non-trivial diophantine equation $x = 0$, whose solution set over $\mathbb{F}_{p^r}$ consists of a single point, one has $N_{p^r} = 1$ for all p and therefore

$$\zeta_p(x = 0; T) = \exp\left(\sum_{r \geq 1} \frac{T^r}{r} \right) = (1 - T)^{-1}. \tag{11}$$

It follows that

$$\zeta(x = 0; s) = \prod_p \left(1 - \frac{1}{p^s}\right)^{-1} \tag{12}$$

$$= \prod_p \left(1 + \frac{1}{p^s} + \frac{1}{p^{2s}} + \frac{1}{p^{3s}} + \cdots\right) \tag{13}$$

$$= \sum_{n=1}^{\infty} \frac{1}{n^s} \quad = \quad \zeta(s). \tag{14}$$

The zeta function of even the humblest diophantine equation is thus a central object of mathematics: the celebrated Riemann zeta function, which is tied to some of the deepest questions concerning the distribution of prime numbers. In his great memoir of 1860, Riemann proved that, even though (13) and (14) only converge absolutely on the right half-plane $\Re(s) > 1$, the function $\zeta(s)$ extends to a meromorphic function of $s \in \mathbb{C}$ (with a single pole at $s = 1$) and possesses an elegant functional equation relating its values at s and $1 - s$. The zeta functions of linear equations $\mathcal{X}$ in $n + 1$ variables are just shifts of the Riemann zeta function, since N_{p^r} is equal to p^{nr}, and therefore $\zeta(\mathcal{X}; s) = \zeta(s - n)$.

Moving on to equations of degree two, the general quadratic equation in one variable is of the form $ax^2 + bx + c = 0$ and its behaviour is governed by its *discriminant*

$$\Delta := b^2 - 4ac. \tag{15}$$

This purely algebraic fact remains true over the finite fields and, for primes $p \nmid 2a\Delta$, one has

$$N_p = \begin{cases} 0 & \text{if } \Delta \text{ is a non-square modulo } p, \\ 2 & \text{if } \Delta \text{ is a square modulo } p. \end{cases} \tag{16}$$

A priori, the criterion for whether $N_p = 2$ or 0 — whether the integer Δ is or is not a quadratic residue modulo p — seems like a subtle condition on the prime p. To get a better feeling for this condition, consider the example of the equation $x^2 - x - 1$, for which $\Delta = 5$. Calculating whether 5 is a square or not modulo p for the first few primes $p \leq 101$ leads to the following list

$$N_p = \begin{cases} 2 & \text{for } p = 11, 19, 29, 31, 41, 59, 61, 71, 79, \\ & \qquad 89, 101, \ldots \\ 0 & \text{for } p = 2, 3, 7, 13, 17, 23, 37, 43, 47, 53, \\ & \qquad 67, 73, 83, \ldots \end{cases} \tag{17}$$

A clear pattern emerges from this experiment: whether $N_p = 0$ or 2 seems to depend only on the rightmost digit of p, i.e. on what the remainder of p is modulo 10. One is led to surmise that

$$N_p = \begin{cases} 2 & \text{if } p \equiv 1, 4 \pmod 5, \\ 0 & \text{if } p \equiv 2, 3 \pmod 5, \end{cases} \tag{18}$$

a formula that represents a dramatic improvement over (16), allowing a much more efficient calculation of N_p for example. The guess in (18) is in fact a consequence of Gauss' celebrated law of quadratic reciprocity:

Theorem (Quadratic reciprocity) *For any equation $ax^2 + bx + c$, with $\Delta := b^2 - 4ac$, the value of the function $p \mapsto N_p$*

(for $p \nmid a\Delta$) depends only on the residue class of p modulo 4Δ, and hence is periodic with period length dividing $4|\Delta|$.

The repeating pattern satisfied by the N_p's as p varies greatly facilitates the manipulation of the zeta functions of quadratic equations. For example, the zeta function of $\mathcal{X} : x^2 - x - 1 = 0$ is equal to

$$\zeta(\mathcal{X}; s) = \zeta(s) \times \Big\{ \Big(1 - \frac{1}{2^s} - \frac{1}{3^s} + \frac{1}{4^s}\Big) + \Big(\frac{1}{6^s} - \frac{1}{7^s} - \frac{1}{8^s} + \frac{1}{9^s}\Big) + \Big(\frac{1}{11^s} - \frac{1}{12^s} - \frac{1}{13^s} + \frac{1}{14^s}\Big) + \cdots \Big\}. \quad (19)$$

The series that occurs on the right side is a prototypical example of a *Dirichlet L-series*. These L-series, which are the key actors in the proof of Dirichlet's theorem on the infinitude of primes in arithmetic progressions, enjoy many of the same analytic properties as the Riemann zeta function: an analytic continuation to the entire complex plane and a functional equation relating their values at s and $1 - s$. They are also expected to satisfy a Riemann hypothesis that generalises Riemann's original statement and is just as deep and elusive.

It is a (not completely trivial) fact that the zeta function of the general quadratic equation in n variables

$$\sum_{i,j=1}^{n} a_{ij} x_i x_j + \sum_{i=1}^{n} b_i x_i + c = 0 \quad (20)$$

involves the same basic constituents – Dirichlet series – as in the one variable case. This means that quadratic diophantine equations in any number of variables are well understood, at least as far as their zeta functions are concerned.

The plot thickens when equations of higher degree are considered. Consider, for instance, the cubic equation $x^3 - x - 1$ of discriminant $\Delta = -23$. For all $p \neq 23$, this cubic equation has no multiple roots over $\mathbb{F}_{p^r}$ and therefore $N_p = 0$, 1 or 3. A simple expression for N_p in this case is given by the following theorem of Hecke:

Theorem (Hecke). *The following holds for all primes $p \neq 23$:*
1. *If p is not a square modulo 23 then $N_p = 1$.*
2. *If p is a square modulo 23 then*

$$N_p = \begin{cases} 0 & \text{if } p = 2a^2 + ab + 3b^2, \\ 3 & \text{if } p = a^2 + ab + 6b^2, \end{cases} \quad (21)$$

for some $a, b \in \mathbb{Z}$.

Hecke's theorem implies that

$$\zeta(x^3 - x - 1; s) \quad = \quad \zeta(s) \times \sum_{n=1}^{\infty} a_n n^{-s}, \quad (22)$$

where the generating series

$$F(q) := \sum_{n=1}^{\infty} a_n q^n = q - q^2 - q^3 + q^6 + q^8 - q^{13} - q^{16} + q^{23} + \cdots \quad (23)$$

is given by the explicit formula

$$F(q) = \frac{1}{2}\left(\sum_{a,b \in \mathbb{Z}} q^{a^2+ab+6b^2} - q^{2a^2+ab+3b^2} \right). \quad (24)$$

The function $f(z) = F(e^{2\pi i z})$ that arises by setting $q = e^{2\pi i z}$ in (24) is a prototypical example of a *modular form*: namely, it satisfies the transformation rule

$$f\left(\frac{az+b}{cz+d}\right) = (cz+d) f(z), \quad \begin{cases} a, b, c, d \in \mathbb{Z}, & ad - bc = 1, \\ 23|c, & \left(\frac{a}{23}\right) = 1, \end{cases} \quad (25)$$

under so-called *modular substitutions* of the form $z \mapsto \frac{az+b}{cz+d}$. This property follows from the *Poisson summation formula* applied to the expression in (24). Thanks to (25), the zeta function of $\mathcal{X}$ can be manipulated with the same ease as the zeta functions of Riemann and Dirichlet. Indeed, Hecke showed that the L-series $\sum_{n=1}^{\infty} a_n n^{-s}$ attached to a modular form $\sum_{n=1}^{\infty} a_n e^{2\pi i n z}$ possess very similar analytic properties, notably an analytic continuation and a Riemann-style functional equation.

The generating series $F(q)$ can also be expressed as an infinite product:

$$\frac{1}{2}\left(\sum_{a,b \in \mathbb{Z}} q^{a^2+ab+6b^2} - q^{2a^2+ab+3b^2} \right) = q \prod_{n=1}^{\infty} (1 - q^n)(1 - q^{23n}). \quad (26)$$

The first few terms of this power series identity can readily be verified numerically but its proof is highly non-obvious and indirect. It exploits the circumstance that the space of holomorphic functions of z satisfying the transformation rule (25) together with suitable growth properties is a one-dimensional complex vector space that also contains the infinite product above. Indeed, the latter is equal to $\eta(q)\eta(q^{23})$, where

$$\eta(q) = q^{1/24} \prod_{n=1}^{\infty} (1 - q^n) \quad (27)$$

is the Dedekind eta function whose logarithmic derivative (after viewing η as a function of z through the change of variables $q = e^{2\pi i z}$) is given by

$$\frac{\eta'(z)}{\eta(z)} = -\pi i \left(\frac{-1}{12} + 2 \sum_{n=1}^{\infty} \Big(\sum_{d|n} d \Big) e^{2\pi i n z} \right) \quad (28)$$

$$= \frac{i}{4\pi} \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \frac{1}{(mz+n)^2}, \quad (29)$$

where the term attached to $(m, n) = (0, 0)$ is excluded from the last sum. The Dedekind η-function is also connected to the generating series for the partition function $p(n)$ describing the number of ways in which n can be expressed as a sum of positive integers via the identity

$$\eta^{-1}(q) = q^{-1/24} \sum_{n=0}^{\infty} p(n) q^n, \quad (30)$$

which plays a starring role alongside Jeremy Irons and Dev Patel in a recent film about the life of Srinivasa Ramanujan.

Commenting on the "unreasonable effectiveness and ubiquity of modular forms", Martin Eichler once wrote: "There are five elementary arithmetical operations: addition, subtraction, multiplication, division, … and modular forms." Equations (26), (29) and (30) are just a few of the many wondrous

identities that abound, like exotic strains of fragrant wild orchids, in what Roger Godement has called the "garden of modular delights".

The example above, and many others of a similar type, are described in Jean-Pierre Serre's delightful monograph [2], touching on themes that were also covered in Serre's lecture at the inaugural Abel Prize ceremony in 2003.

Hecke was able to establish that all cubic polynomials in one variable are *modular*, i.e., the coefficients of their zeta functions obey patterns just like those of (24) and (25). Wiles' achievement was to extend this result to a large class of cubic diophantine equations in two variables over the rational numbers: the *elliptic curve* equations, which can be brought into the form

$$y^2 = x^3 + ax + b \tag{31}$$

after a suitable change of variables and which are non-singular, a condition equivalent to the assertion that the *discriminant* $\Delta := -16(4a^3 + 27b^2)$ is non-zero.

To illustrate Wiles' theorem with a concrete example, consider the equation

$$E : y^2 = x^3 - x, \tag{32}$$

of discriminant $\Delta = 64$. After setting

$$\zeta(E; s) = \zeta(s-1) \times (a_1 + a_2 2^{-s} + a_3 3^{-s} + a_4 4^{-s} + \cdots)^{-1}, \tag{33}$$

the associated generating series satisfies the following identities reminiscent of (24) and (26),

$$\begin{aligned} F(q) &= \sum a_n q^n \\ &= q - 2q^5 - 3q^9 + 6q^{13} + 2q^{17} - q^{25} + \cdots \qquad (34) \\ &= \sum_{a,b} a \cdot q^{(a^2+b^2)} \qquad (35) \\ &= q \prod_{n=1}^{\infty} (1-q^{4n})^2 (1-q^{8n})^2, \qquad (36) \end{aligned}$$

where the sum in (35) runs over the $(a, b) \in \mathbb{Z}^2$ for which the Gaussian integer $a + bi$ is congruent to 1 modulo $(1 + i)^3$. (This identity follows from Deuring's study of zeta functions of elliptic curves *with complex multiplication*, and may even have been known earlier.) Once again, the holomorphic function $f(z) := F(e^{2\pi iz})$ is a modular form, satisfying the slightly different transformation rule

$$f\left(\frac{az+b}{cz+d}\right) = (cz+d)^2 f(z), \quad \begin{cases} a, b, c, d \in \mathbb{Z}, & ad - bc = 1, \\ 32|c. \end{cases} \tag{37}$$

Note the exponent 2 that appears in this formula. Because of it, the function $f(z)$ is said to be a *modular form of weight* 2 and level 32. The modular forms of (25) attached to cubic equations in one variable are of weight 1 but otherwise the parallel of (35) and (36) with (24) and (26) is striking. The original conjecture of Shimura-Taniyama and Weil asserts the same pattern for all elliptic curves:

Conjecture (Shimura, Taniyama, Weil). *Let E be any elliptic curve. Then,*

$$\zeta(E; s) = \zeta(s-1) \times \left(\sum_{n=1}^{\infty} a_n n^{-s} \right)^{-1}, \tag{38}$$

where $f_E(z) := \sum a_n e^{2\pi inz}$ is a modular form of weight 2.

The conjecture was actually more precise and predicted that the level of f_E – i.e., the integer that appears in the transformation property for f_E, as the integers 23 and 32 do in (25) and (37) respectively – is equal to the *arithmetic conductor* of E. This conductor, which is divisible only by primes for which the equation defining E becomes singular modulo p, is a measure of the arithmetic complexity of E and can be calculated explicitly from an equation for E by an algorithm of Tate. An elliptic curve is said to be *semistable* if its arithmetic conductor is squarefree. This class of elliptic curves includes those of the form

$$y^2 = x(x-a)(x-b), \tag{39}$$

with $\gcd(a, b) = 1$ and $16|b$. The most famous elliptic curves in this class are those that ultimately do not exist: the "Frey curves" $y^2 = x(x - a^p)(x + b^p)$ arising from putative solutions to Fermat's equation $a^p + b^p = c^p$, whose non-existence had previously been established in a landmark article of Kenneth Ribet, under the assumption of their modularity. It is the proof of the Shimura-Taniyama-Weil conjecture for semistable elliptic curves that earned Andrew Wiles the Abel prize:

Theorem (Wiles). *Let E be a semistable elliptic curve. Then E satisfies the Shimura-Taniyama-Weil conjecture.*

The semistability assumption in Wiles' theorem was later removed by Christophe Breuil, Brian Conrad, Fred Diamond and Richard Taylor around 1999. (See, for instance, the account that appeared in the Notices at the time [1].)

As a prelude to describing some of the important ideas in its proof, one must first try to explain why Wiles' theorem occupies such a central position in mathematics. The Langlands programme places it in a larger context by offering a vast generalisation of what it means for a diophantine equation to be "associated to a modular form". The key is to view modular forms attached to cubic equations or to elliptic curves, as in (24) or (34), as vectors in certain irreducible infinite-dimensional representations of the locally compact topological group

$$\mathbf{GL}_2(\mathbb{A}_{\mathbb{Q}}) = \prod_p{}' \mathbf{GL}_2(\mathbb{Q}_p) \times \mathbf{GL}_2(\mathbb{R}), \tag{40}$$

where $\prod'_p$ denotes a restricted direct product over all the prime numbers, consisting of elements $(\gamma_p)_p$ for which the p-th component γ_p belongs to the maximal compact subgroup $\mathbf{GL}_2(\mathbb{Z}_p)$ for all but finitely many p. The shift in emphasis from modular forms to the so-called *automorphic representations* that they span is decisive. Langlands showed how to attach an L-function to any irreducible automorphic representation of $G(\mathbb{A}_{\mathbb{Q}})$ for an arbitrary reductive algebraic group G, of which the matrix groups $\mathbf{GL}_n$ and more general algebraic

groups of Lie type are prototypical examples. This greatly enlarges the notion of what it means to be "modular": a diophantine equation is now said to have this property if its zeta function can be expressed in terms of the Langlands L-functions attached to automorphic representations. One of the fundamental goals in the Langlands programme is to establish further cases of the following conjecture:

Conjecture. *All diophantine equations are modular in the above sense.*

This conjecture can be viewed as a far-reaching generalisation of quadratic reciprocity and underlies the non-abelian reciprocity laws that are at the heart of Andrew Wiles' achievement.

Before Wiles' proof, the following general classes of diophantine equations were known to be modular:

- Quadratic equations, by Gauss' law of quadratic reciprocity.
- Cubic equations in one variable, by the work of Hecke and Maass.
- Quartic equations in one variable.

This last case deserves further comment, since it has not been discussed previously and plays a primordial role in Wiles' proof. The modularity of quartic equations follows from the seminal work of Langlands and Tunnell. While it is beyond the scope of this survey to describe their methods, it must be emphasised that Langlands and Tunnell make essential use of the *solvability by radicals* of the general quartic equation, whose underlying symmetry group is contained in the permutation group S_4 on 4 letters. Solvable extensions are obtained from a succession of abelian extensions, which fall within the purview of class field theory developed in the 19th and first half of the 20th centuries. On the other hand, the modularity of the general equation of degree > 4 in one variable, which cannot be solved by radicals, seemed to lie well beyond the scope of the techniques that were available in the "pre-Wiles era". The reader who perseveres to the end of this essay will be given a glimpse of how our knowledge of the modularity of the general quintic equation has progressed dramatically in the wake of Wiles' breakthrough.

Prior to Wiles' proof, modularity was also not known for any interesting general class of equations (of degree > 2, say) in more than one variable; in particular, it had only been verified for finitely many elliptic curves over $\mathbb{Q}$ up to isomorphism over $\bar{\mathbb{Q}}$ (including the elliptic curves over $\mathbb{Q}$ with complex multiplication, of which the elliptic curve of (31) is an example). Wiles' modularity theorem confirmed the Langlands conjectures in the important test case of elliptic curves, which may seem to be (and, in fact, are) very special diophantine equations but which have provided a fertile terrain for arithmetic investigations, both in theory and in applications (e.g., cryptography and coding theory).

Returning to the main theme of this report, Wiles' proof is also important for having introduced a revolutionary new approach, which has opened the floodgates for many further breakthroughs in the Langlands programme.

To expand on this point, we need to present another of the *dramatis personae* in Wiles' proof: *Galois representations*. Let $G_{\mathbb{Q}} = \mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q})$ be the absolute Galois group of $\mathbb{Q}$, namely, the automorphism group of the field of all algebraic numbers. It is a profinite group, endowed with a natural topology for which the subgroups $\mathrm{Gal}(\bar{\mathbb{Q}}/L)$ with L ranging over the finite extensions of $\mathbb{Q}$ form a basis of open subgroups. Following the original point of view taken by Galois himself, the group $G_{\mathbb{Q}}$ acts naturally as permutations on the roots of polynomials with rational coefficients. Given a finite set S of primes, one may consider only the monic polynomials with integer coefficients whose discriminant is divisible only by primes $\ell \in S$ (eventually after a change of variables). The topological group $G_{\mathbb{Q}}$ operates on the roots of such polynomials through a quotient, denoted $G_{\mathbb{Q},S}$: the automorphism group of the maximal algebraic extension *unramified* outside of S, which can be regarded as the symmetry group of all the zero-dimensional varieties over $\mathbb{Q}$ having "non-singular reduction outside of S".

In addition to the permutation representations of $G_{\mathbb{Q}}$ that were so essential in Galois' original formulation of his theory, it has become important to study the (continuous) *linear* representations

$$\varrho : G_{\mathbb{Q},S} \longrightarrow GL_n(L) \tag{41}$$

of this Galois group, where L is a complete field, such as the fields $\mathbb{R}$ or $\mathbb{C}$ of real or complex numbers, the finite field $\mathbb{F}_{\ell^r}$ equipped with the discrete topology, or a finite extension $L \subset \bar{\mathbb{Q}}_\ell$ of the field $\mathbb{Q}_\ell$ of ℓ-adic numbers.

Galois representations were an important theme in the work of Abel and remain central in modern times. Many illustrious mathematicians in the 20th century have contributed to their study, including three former Abel prize winners: Jean-Pierre Serre, John Tate and Pierre Deligne. Working on Galois representations might seem to be a prerequisite for an algebraic number theorist to receive the Abel prize!

Like diophantine equations, Galois representations also give rise to analogous zeta functions. More precisely, the group $G_{\mathbb{Q},S}$ contains, for each prime $p \notin S$, a distinguished element called the *Frobenius element* at p, denoted σ_p. Strictly speaking, this element is only well defined up to conjugacy in $G_{\mathbb{Q},S}$ but this is enough to make the arithmetic sequence

$$N_{p^r}(\varrho) := \mathrm{Trace}(\varrho(\sigma_p^r)) \tag{42}$$

well defined. The zeta function $\zeta(\varrho; s)$ packages the information from this sequence in exactly the same way as in the definition of $\zeta(\mathcal{X}; s)$.

For example, if $\mathcal{X}$ is attached to a polynomial P of degree d in one variable, the action of $G_{\mathbb{Q},S}$ on the roots of P gives rise to a d-dimensional permutation representation

$$\varrho_{\mathcal{X}} : G_{\mathbb{Q},S} \longrightarrow \mathbf{GL}_d(\mathbb{Q}) \tag{43}$$

and $\zeta(\mathcal{X}, s) = \zeta(\varrho_{\mathcal{X}}, s)$. This connection goes far deeper, extending to diophantine equations in $n+1$ variables for general $n \geq 0$. The glorious insight at the origin of the Weil conjectures is that $\zeta(\mathcal{X}; s)$ can be expressed in terms of the zeta functions of Galois representations arising in the *étale cohomology* of $\mathcal{X}$, a cohomology theory with ℓ-adic coefficients that associates to $\mathcal{X}$ a collection

$$\left\{H^i_{\mathrm{et}}(\mathcal{X}_{/\bar{\mathbb{Q}}}, \mathbb{Q}_\ell)\right\}_{0 \leq i \leq 2n}$$

of finite-dimensional $\mathbb{Q}_\ell$-vector spaces endowed with a continuous linear action of $G_{\mathbb{Q},S}$. (Here, S is the set of primes q consisting of ℓ and the primes for which the equation of $\mathcal{X}$ becomes singular after being reduced modulo q.) These representations generalise the representation $\varrho_{\mathcal{X}}$ of (43), insofar as the latter is realised by the action of $G_{\mathbb{Q},S}$ on $H^0_{\text{et}}(\mathcal{X}_{\bar{\mathbb{Q}}}, \mathbb{Q}_\ell)$ after extending the coefficients from $\mathbb{Q}$ to $\mathbb{Q}_\ell$.

Theorem (Weil, Grothendieck, ...). *If $\mathcal{X}$ is a diophantine equation having good reduction outside of S, there exist Galois representations ϱ_1 and ϱ_2 of $G_{\mathbb{Q},S}$ for which*

$$\zeta(\mathcal{X}; s) = \zeta(\varrho_1; s)/\zeta(\varrho_2; s). \tag{44}$$

The representations ϱ_1 and ϱ_2 occur in $\oplus H^i_{\text{et}}(\mathcal{X}_{/\bar{\mathbb{Q}}}, \mathbb{Q}_\ell)$, where the direct sum ranges over the odd and even values of $0 \leq i \leq 2n$ for ϱ_1 and ϱ_2 respectively. More canonically, there are always *irreducible* representations $\varrho_1, \ldots, \varrho_t$ of $G_{\mathbb{Q},s}$ and integers $d_1, \ldots d_t$ such that

$$\zeta(\mathcal{X}; s) = \prod_{i=1}^{t} \zeta(\varrho_i; s)^{d_i}, \tag{45}$$

arising from the decompositions of the (semisimplification of) the $H^i_{\text{et}}(\mathcal{X}_{\bar{\mathbb{Q}}}, \mathbb{Q}_\ell)$ into a sum of irreducible representations. The $\zeta(\varrho_i, s)$ can be viewed as the "atomic constituents" of $\zeta(\mathcal{X}, s)$, and reveal much of the "hidden structure" in the underlying equation. The decomposition of $\zeta(\mathcal{X}; s)$ into a product of different $\zeta(\varrho_i; s)$ is not unlike the decomposition of a wave function into its simple harmonics.

A Galois representation is said to be *modular* if its zeta function can be expressed in terms of generating series attached to modular forms and automorphic representations, and is said to be *geometric* if it can be realised in an étale cohomology group of a diophantine equation as above. The "main conjecture of the Langlands programme" can now be amended as follows:

Conjecture. *All geometric Galois representations of $G_{\mathbb{Q},S}$ are modular.*

Given a Galois representation

$$\varrho : G_{\mathbb{Q},S} \longrightarrow \mathbf{GL}_n(\mathbb{Z}_\ell) \tag{46}$$

with ℓ-adic coefficients, one may consider the resulting mod ℓ representation

$$\bar{\varrho} : G_{\mathbb{Q},S} \longrightarrow \mathbf{GL}_n(\mathbb{F}_\ell). \tag{47}$$

The passage from ϱ to $\bar{\varrho}$ amounts to replacing the quantities $N_{p^r}(\varrho) \in \mathbb{Z}_\ell$ as p^r ranges over all the prime powers with their mod ℓ reduction. Such a passage would seem rather contrived for the sequences $N_{p^r}(\mathcal{X})$ – why study the solution counts of a diophantine equation over different finite fields, taken modulo ℓ? – if one did not know *a priori* that these counts arise from ℓ-adic Galois representations with coefficients in $\mathbb{Z}_\ell$. There is a corresponding notion of what it means for $\bar{\varrho}$ to be modular, namely, that the data of $N_{p^r}(\bar{\varrho})$ agrees, very loosely speaking, with the mod ℓ reduction of similar data arising from an automorphic representation. We can now state Wiles' celebrated *modularity lifting theorem*, which lies at the heart of his strategy:

Wiles' modularity lifting theorem. *Let*

$$\varrho : G_{\mathbb{Q},S} \longrightarrow \mathbf{GL}_2(\mathbb{Z}_\ell) \tag{48}$$

be an irreducible geometric Galois representation satisfying a few technical conditions (involving, for the most part, the restriction of ϱ to the subgroup $G_{\mathbb{Q}_\ell} = Gal(\bar{\mathbb{Q}}_\ell/\mathbb{Q}_\ell)$ of $G_{\mathbb{Q},S}$). If $\bar{\varrho}$ is modular and irreducible then so is ϱ.

This stunning result was completely new at the time: nothing remotely like it had ever been proved before! Since then, "modularity lifting theorems" have proliferated and their study, in ever more general and delicate settings, has spawned an industry and led to a plethora of fundamental advances in the Langlands programme.

Let us first explain how Wiles himself parlays his original modularity lifting theorem into a proof of the Shimura-Taniyama-Weil conjecture for semistable elliptic curves. Given such an elliptic curve E, consider the groups

$$E[3^n] := \left\{P \in E(\bar{\mathbb{Q}}) \ : \ 3^n P = 0\right\}, \qquad T_3(E) := \lim_{\leftarrow} E[3^n], \tag{49}$$

the inverse limit being taken relative to the multiplication-by-3 maps. The groups $E[3^n]$ and $T_3(E)$ are free modules of rank 2 over $(\mathbb{Z}/3^n\mathbb{Z})$ and $\mathbb{Z}_3$ respectively and are endowed with continuous linear actions of $G_{\mathbb{Q},S}$, where S is a set of primes containing 3 and the primes that divide the conductor of E. One obtains the associated Galois representations:

$$\begin{aligned} \bar{\varrho}_{E,3} &: G_{\mathbb{Q},S} \longrightarrow \text{Aut}(E[3]) \simeq \mathbf{GL}_2(\mathbb{F}_3), \\ \varrho_{E,3} &: G_{\mathbb{Q},S} \longrightarrow \mathbf{GL}_2(\mathbb{Z}_3). \end{aligned} \tag{50}$$

The theorem of Langlands and Tunnell about the modularity of the general quartic equation leads to the conclusion that $\bar{\varrho}_{E,3}$ is modular. This rests on the happy circumstance that

$$\mathbf{GL}_2(\mathbb{F}_3)/\langle \pm 1 \rangle \simeq S_4 \tag{51}$$

and, hence, that $E[3]$ has essentially the same symmetry group as the general quartic equation! The isomorphism in (51) can be realised by considering the action of $\mathbf{GL}_2(\mathbb{F}_3)$ on the set $\{0, 1, 2, \infty\}$ of points on the projective line over $\mathbb{F}_3$.

If E is semistable, Wiles is able to check that both $\varrho_{E,3}$ and $\bar{\varrho}_{E,3}$ satisfy the conditions necessary to apply the modularity lifting theorem, at least when $\bar{\varrho}_{E,3}$ is *irreducible*. It then follows that $\varrho_{E,3}$ is modular and therefore so is E, since $\zeta(E; s)$ and $\zeta(\varrho_{E,3}; s)$ are the same.

Note the key role played by the result of Langlands-Tunnell in the above strategy. It is a dramatic illustration of the unity and historical continuity of mathematics that the solution in radicals of the general quartic equation, one of the great feats of the algebraists of the Italian renaissance, is precisely what allowed Langlands, Tunnell and Wiles to prove their modularity results more than five centuries later.

Having established the modularity of all semistable elliptic curves E for which $\bar{\varrho}_{E,3}$ is irreducible, Wiles disposes of the others by applying his lifting theorem to the prime $\ell = 5$ instead of $\ell = 3$. The Galois representation $\bar{\varrho}_{E,5}$ is always irreducible in this setting because no elliptic curve over $\mathbb{Q}$ can

have a rational subgroup of order 15. Nonetheless, the approach of exploiting $\ell = 5$ seems hopeless at first glance because the Galois representation $E[5]$ is not known to be modular a priori, for much the same reason that the general quintic equation cannot be solved by radicals. (Indeed, the symmetry group $\mathbf{SL}_2(\mathbb{F}_5)$ is a double cover of the alternating group A_5 on 5 letters and thus closely related to the symmetry group underlying the general quintic.) To establish the modularity of $E[5]$, Wiles constructs an auxiliary semistable elliptic curve E' satisfying

$$\bar{\varrho}_{E',5} = \bar{\varrho}_{E,5}, \qquad \bar{\varrho}_{E',3} \text{ is irreducible.} \tag{52}$$

It then follows from the argument in the previous paragraph that E' is modular, hence that $E'[5] = E[5]$ is modular as well, putting E within striking range of the modularity lifting theorem with $\ell = 5$. This lovely epilogue of Wiles' proof, which came to be known as the "3-5 switch", may have been viewed as an expedient trick at the time. But, since then, the prime switching argument has become firmly embedded in the subject and many variants of it have been exploited to spectacular effect in deriving new modularity results.

Wiles' modularity lifting theorem reveals that "modularity is contagious" and can often be passed on to an ℓ-adic Galois representation from its mod ℓ reduction. It is this simple principle that accounts for the tremendous impact that the modularity lifting theorem and the many variants proven since then continue to have on the subject. Indeed, the modularity of elliptic curves was only the first in a series of spectacular applications of the ideas introduced by Wiles and, since 1994, the subject has witnessed a real golden age, in which open problems that previously seemed completely out of reach have succumbed one by one.

Among these developments, let us mention a few below:

- The two-dimensional Artin conjecture, first formulated in 1923, concerns the modularity of all odd, two-dimensional Galois representations

$$\varrho : G_{\mathbb{Q},S} \longrightarrow \mathbf{GL}_2(\mathbb{C}). \tag{53}$$

  The image of such a ϱ modulo the scalar matrices is isomorphic either to a dihedral group, to A_4, to S_4 or to A_5. Thanks to the earlier work of Hecke, Langlands and Tunnell, only the case of the projective image A_5 remained to be disposed of. Many new cases of the two-dimensional Artin conjecture were proven in this setting by Kevin Buzzard, Mark Dickinson, Nick Shepherd-Barron and Richard Taylor around 2003, using the modularity of all mod 5 Galois representations arising from elliptic curves as a starting point.
- Serre's Conjecture, which was formulated in 1987, asserts the modularity of all odd, two-dimensional Galois representations

$$\varrho : G_{\mathbb{Q},S} \longrightarrow \mathbf{GL}_2(\mathbb{F}_{p^r}), \tag{54}$$

  with coefficients in a finite field. This result was proven by Chandrasekhar Khare and Jean-Pierre Wintenberger in 2008 using a glorious extension of the "3 – 5 switching technique" in which essentially all the primes are used. (See Khare's report in the Notices of the AMS mentioned above.) This result also implies the two-dimensional Artin conjecture in the general case.
- The two-dimensional Fontaine–Mazur conjecture concerning the modularity of odd, two-dimensional p-adic Galois representations

$$\varrho : G_{\mathbb{Q},S} \longrightarrow \mathbf{GL}_2(\bar{\mathbb{Q}}_p) \tag{55}$$

  satisfying certain technical conditions with respect to their restrictions to the Galois group of $\mathbb{Q}_p$ was proven in many cases as a consequence of work of Pierre Colmez, Matthew Emerton and Mark Kisin.
- The Sato–Tate conjecture concerning the distribution of the numbers $N_p(E)$ for an elliptic curve E as the prime p varies, whose proof was known to follow from the modularity of all the symmetric power Galois representations attached to E, was proven in large part by Laurent Clozel, Michael Harris, Nick Shepherd-Barron and Richard Taylor around 2006.
- One can also make sense of what it should mean for diophantine equations over more general number fields to be modular. The modularity of elliptic curves over all real quadratic fields has been proven very recently by Nuno Freitas, Bao Le Hung and Samir Siksek by combining the ever more general and powerful modularity lifting theorems currently available with a careful diophantine study of the elliptic curves that could a priori fall outside the scope of these lifting theorems.
- Among the spectacular recent developments building on Wiles' ideas is the proof, by Laurent Clozel and Jack Thorne, of the modularity of certain symmetric powers of the Galois representations attached to holomorphic modular forms, which is described in Thorne's contribution to the Notices of the AMS mentioned above.

These results are just a sample of the transformative impact of modularity lifting theorems. The Langlands programme remains a lively area, with many alluring mysteries yet to be explored. It is hard to predict where the next breakthroughs will come but surely they will continue to capitalise on the rich legacy of Andrew Wiles' marvellous proof.

References

[1] H. Darmon. *A proof of the full Shimura-Taniyama-Weil conjecture is announced*. Notices of the AMS **46** (1999) no. 11, 1397–1401.

[2] J-P. Serre. Lectures on $N_X(p)$. Chapman & Hall/CRC Research Notes in Mathematics, 11. CRC Press, Boca Raton, FL, 2012. x+163 pp.

Henri Darmon [darmon@math.mcgill.ca] is a James McGill Professor of Mathematics at McGill University and a member of CICMA (Centre Interuniversitaire en Calcul Mathématique Algébrique) and CRM (Centre de Recherches Mathématiques). He received the 2017 AMS Cole Prize in Number Theory and the 2017 CRM-Fields-PIMS Prize for his contributions to the arithmetic of elliptic curves and modular forms.